**Literature Review**

Bayesian Changepoint Detection in Textual Data Streams

Student ID: 22999825

Name: Andisheh Partovi

Course: Honours Degree of Bachelor of Computer Science

Date: 08/09/2014

Supervisor: Gholamreza Haffari – Ingrid Zukerman

# Abstract

As web social media can provide vast amounts of user generated content that can reflect the thoughts and opinions of the users, it's an invaluable source for sentiment analysis research and therefore has been extensively used to extract a variety of information. But temporal analysis is often missing from sentiment analysis studies. Performing temporal analysis on social media can provide information not only regarding opinions and sentiments, but also regarding the changes in them over time. In this research, a real-time changepoint detection model will be developed that can be applied to streams of textual input from Twitter in order to reveal interesting trends in them over time. In this literature review, I provide an overview of different changepoint detection techniques along with previous work on sentiment analysis over social media.

Contents

[Abstract 2](#_Toc403308639)

[1. Introduction 4](#_Toc403308640)

[2. Sentiment Analysis on Twitter 5](#_Toc403308641)

[2.1. Twitter Anatomy 5](#_Toc403308642)

[2.2. Previous Work on Sentiment Analysis 7](#_Toc403308643)

[3. The Changepoint Detection Problem 8](#_Toc403308644)

[3.1. The Existing Methods 9](#_Toc403308645)

[3.3.1 MCMC Methods 9](#_Toc403308646)

[3.3.2 Message Passing Methods 10](#_Toc403308647)

[3.3.3 Other Bayesian Approaches 12](#_Toc403308648)

[4. Changepoint Detection on Twitter 13](#_Toc403308649)

[References: 15](#_Toc403308650)

# 1. Introduction

The task of analysing opinions from text data is known as “sentiment analysis” or “opinion mining” (Pang and Lee, 2008) and is used to extract subjective information from textual source material. As web social media, blogging, and online forums can provide vast amounts of user generated content that can reflect the thoughts and opinions of the users, it is an invaluable source for any opinion mining applications, and hence has been studied extensively in recent years.

Sentiment analysis on social media is not restricted to Twitter. Thelwall and Prabowo (2007) as well as Bansal and Koudas (2007) worked on online blogs, Kramer (2010) analysed Facebook status updates to estimate national happiness, and Kennedy and Inkpen (2006) as well as Pang et al. (2002) worked on classifying online movie reviews. However, since its launch in 2006, Twitter has attracted more and more researchers.

As of November 2014, people post more than 500 million Twitter messages (called tweets) every day (About.twitter.com, 2014), yielding a noisy but sometimes informative corpus of 140-character messages that reflects their daily activities and thoughts as well as current events in an unprecedented manner (Ritter et al., 2011).

Sentiment analysis on Twitter has been carried out to extract a variety of different information. In one study, Bollen et al. (2009) analysed the six dimensions of emotion in Twitter, showing that these typically reflect significant offline events. In another study, Bollen et al. (2011) correlated Twitter mood to the changes in the stock market. Jansen et al. (2009) used tweets to automatically extract customer opinions about products or brands; Lampos and Cristianini (2010), Lampos et al. (2010), Paul and Dredze (2011), and Collier and Doan (2011) used tweets to track infectious diseases; Sakaki et al. (2010) used tweets to detect earthquakes; and Lampos et al. (2013) used tweets to predict election winners.

Despite all the research in this field, one of the areas that has been left relatively untouched is applying temporal analysis to tweets. If the stream of tweets is analysed over time, interesting trends such as changes in topics of interest, meanings of words, or sentiments over time can be revealed. In other words, performing temporal analysis on social media can provide information not only regarding opinions, but also regarding the changes in opinions over time.

One of the oldest statistical tools that has been utilised in many problem domains (see Section 3) and can be employed in this problem is *changepoint detection*. In this research, a changepoint detection model will be developed that can be applied to streams of textual data (Twitter messages) in order to reveal interesting trends in them over time.

In this literature review, I will provide a brief description of Twitter, its unique features, and some of the previous sentiment analysis research on it (Section 2); followed by introducing the changepoint detection problem, its applications, and an analysis of some of the existing methods (Section 3); and reviewing the few existing studies on applying changepoint detection techniques to Twitter (Section 4).

# 2. Sentiment Analysis on Twitter

## 2.1. Twitter Anatomy

Twitter is an online social networking service that enables users to send and read short 140-character messages. Twitter is mainly used by the public to share information and to describe minor daily activities (Java et al., 2007). Although it can also be used for information dissemination, for example, by government organizations (Wigand, 2010) or private companies. About 80% of Twitter users update followers on what they are currently doing, while the remainder centre on information (Naaman et al., 2010).

What is Twitter?

Tweets are short because of the 140-character limit (117 if they contain a URL) and therefore have undesirable qualities such as extensive presence of chat acronyms (such as FTW for "for the win" or OMG for "oh my God") in addition to qualities common to most online media content such as use of colloquial terms, emoticons (like ☺) and emoji[[1]](#footnote-1), spelling errors, and alternative spellings to emphasise a sentiment (such as “reallyyyyy” for “really”).

Because of these features, the performance of standard Natural Language Processing (NLP) tools can be severely degraded on tweets (Ritter et al., 2011). When analysing these data, one approach is to take them as they are, without any modifications. Conversely, the mentioned properties can be eliminated through normalisation techniques such as designing specialised dictionaries for emoticons and acronyms and substituting regular expressions resembling a word with that word. Additionally, standard natural language preprocessing techniques such as decapitalisation and stop word removal might be necessary.

Pre-processing

Another preprocessing step that might be necessary for twitter is removing tweets with a URL, which are most likely spam (Lim and Buntine, 2014). This is a conservative approach taken by most researchers that might result in loss of information a not all tweets containing URL are spam. One particular study (Wu et al., 2011) focused on URLs in the tweets and studied their popularity lifespan.

Two prominent features that used to be unique to tweets, but are now used in all forms of online communications, are hashtags[[2]](#footnote-2) and mentions[[3]](#footnote-3).

Hashtags were invented by Twitter users in early 2008 and have emerged as a method for filtering and promoting content in Twitter, rather than as a tool for retrieval (Huang et al., 2010). Hashtags are informal since they have no standards and can be used as either inline words or categorical labels.

Hashtags

Hashtags can be strong indicators of topics for tweets (Mehrotra et al., 2013) and therefore have been used as a sentiment analysis tool in previous work. Romero et al. (2011) and Kwak et al. (2010) used them for topic identification. Preotiuc-Pietro and Cohn (2013) have studied hashtag distributions in order to aid the classification of tweets based on their topics, and successfully improved the performance of their Naïve Bayes Classifier by providing a better prior knowledge of hashtags. Kunneman et al. (2014) attempted to reassign hashtags to tweets that were stripped from their original hashtags, and evaluated the system using the original hashtags.

Sentiment Analysis Using Hashtags

About 31% of Tweets seem to be directed at a specific user using mentions (Boyd et al., 2009), emphasising the social element of Twitter and its usage as a chatting system rather than an information broadcasting system. Takahashi et al. (2011) proposed a probability model of the mentioning behaviour as part of their study on topic emergence in Twitter.

Mentions

## 2.2. Previous Work on Sentiment Analysis

Numerous methods have been applied to address different problems in sentiment analysis over social media and there is a large volume of literature covering this area. In this section, I focus on the few studies that considered the element of time in their sentiment analysis in order to highlight the gap in this field.

Most of the existing temporal analysis literature focuses on topic detection and tracking (TDT) where temporal patterns associated with tweet content are studied, such as how the content’s popularity grows and fades over time. For instance, Yang and Leskovec (2011) performed K-Spectral Centroid (K-SC) clustering on topic time-series extracted from tweets in order to uncover the temporal dynamics of the content. Cataldi et al. (2010) proposed a topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community.

Temporal Analysis of Tweets

These studies are sometimes accompanied by a spatial analysis of the users by utilising graph analysis techniques on a follower-following network (FFN). Kwak et.al. (2010) and Ardon et al. (2011) studied several aspects of topic diffusion and information propagation in the FFN. Their temporal analysis of trending topics on Twitter, however, was limited to plotting the topic change over time. The focus of the latter study was mostly on identifying topic initiators, and how topics spread inside the network.

Temporospatial Analysis

Some researchers focused on temporal analysis of other aspects of Twitter. For example, Abel et al. (2011), as part of their user modelling study, conducted a temporal analysis of Twitter user profiles, for example, they examined whether profiles generated on the weekends differ from those generated during the week. Huang et al. (2010) further characterised the temporal dynamics of hash-tags via statistical measures such as standard deviation and kurtosis. They discovered that some hashtags are widely used for a few days but then disappear quickly. Wu et al. (2011a and 2011b) studied the temporal dynamics of the URL links in tweets and estimated their popularity life span.

Other Temporal Analysis

# 3. The Changepoint Detection Problem

Identifying abrupt changes in a stream of data, called *changepoint detection*, has proven to be useful in many problem domains and hence has occupied the minds of researchers in the statistics and data mining communities for years.

One of the early applications of changepoint detection was quality control and production monitoring where decisions are to be reached regarding the quality of the products or their classification in real time when their measurements are taken. This process might require fast decision making when the safety of employees is involved, so quick and accurate detection of abrupt changes becomes essential (Basseville and Nikiforov, 1993).

Changepoint detection is often studied in association with time-series. Time-series is an ordered sequence of data points. The ordering of the data points is mostly through time, particularly equally spaced time intervals. The number of monthly airline passengers in the US, or the US dollar to Euro daily exchange rate are two examples of time-series (Madsen, 2007).

Time-series

Changepoints may represent important events in the time-series and can partition it into independent segments. Recognition-oriented signal processing benefits from the segmentation provided by changepoint detection approaches, and therefore has been used in processing a range of signals, including biomedical signals such as EEGs (Bodenstein and Praetorius, 1977; Barlow et al., 1981).

Applications

In addition to the mentioned applications, changepoint detection has been utilised in a myriad of other problem domains, examples of which include detecting changes and possibly predicting them in stock markets (Koop and Potter, 2004; Xuan and Murphy, 2007), understanding climate change (Reeves et al., 2007; Beaulieu et al., 2012), genetics and DNA segmentation (Wang et al., 2011; Fearnhead and Liu, 2007), disease demographics (Dension and Holmes, 2001), intrusion detection in computer networks (Yamanishi et al., 2000), satellite imagery analysis (Bovolo et al., 2008; Habib et al., 2009), and even detecting changes in animal behaviours (Roth et al., 2012).

Based on the detection delay, changepoint detection methods can be categorised as online (real-time) or offline (retrospective) detections. Online detection analyses the data stream as it becomes available, and is utilised in problems that demand immediate responses like a robot’s navigational system that has to react to a dynamically changing environment. Offline detection, which comprises most of the research in this field, uses the entire dataset to identify the changepoint locations, and is applied to the problems that can afford computational delays. Any offline problem can also be approached by online methods, by introducing a time for each observation, but not vice versa.

Online Vs. offline Detection

## 3.1. The Existing Methods

The changepoint detection problem has been studied for decades and a large number of methods have been proposed to address it in different problem domains (Basseville and Nikiforov, 1993; Brodsky and Darkhovsky, 1993; Csorgo and Horvath, 1997; Chen and Gupta 2000; Gustafsson, 2000). In this section, an overview of some of the Bayesian changepoint detection methods reviewed for this research is provided, along with some analysis of their relative merits and disadvantages. A complete review of all the changepoint detection literature is infeasible considering the volume of it, which according to Carlin et al. (1992), as of 1992, was enormous.

In Bayesian approaches a prior distribution over the number and location of changepoints is assumed, and Bayesian inference to calculate the posterior distribution is performed. Exact computation of the posterior distribution over changepoint configurations is intractable for large data sets. Therefore, different techniques are employed to do an approximate inference.

Bayesian Change Detection

### 3.3.1 MCMC Methods

Markov chain Monte Carlo (MCMC) are a large class of sampling algorithms that are often applied to solve integration and optimisation problems in high-dimensional spaces. These algorithms have played a significant role in many areas of science and engineering over the last two decades (Andrieu et al., 2003).

Using MCMC algorithms for posterior sampling in changepoint models has been studied as an offline changepoint detection technique for years. Carlin et al. (1992) devised a Gibbs sampler (Geman and Geman, 1984) for Bayesian changepoint models where the number of changepoints was known to be one. This method was later extended to multiple changepoints models by Stephens (1994) and Chib (1998). It is known that Gibbs samplers can suffer from very slow convergences (Whiteley et al., 2011) and moreover requires a knowledge of the number of changepoints. Hence, other algorithms were devised to address MCMC methods’ shortcomings.

Gibbs Sampler

Reversible-jump MCMC sampling introduced by Green (1995) works even if the number of parameters in the model (here the number of changepoints) is unknown or changes over time but at the price of an even slower convergence. Therefore, it is still not efficient enough for online changepoint detection.

Reversible-jump MCMC

### 3.3.2 Message Passing Methods

Fearnhead and Liu (2007), as well as Adams and MacKay (2007), have independently worked on developing message passing algorithms efficient enough to calculate the posterior probability distribution of the changepoints in real time. Given the superiority of these two online approaches and their successful deployment in different problem domains, we have chosen to use them as the basis of our changepoint detection model.

Their models are largely based on the "Product Partition" model introduced by Barry and Hartigan (1992). This model assumes that time-series data can be partitioned into independent and identically distributed (i.i.d.) partitions, separated by the points where the data’s generative parameters change; i.e. given a set of observations collected over time, these models introduce a number of changepoints which split the data into a set of disjoint segments. It is then assumed that the data arise from a single model within each segment, but with different models across the segments.

Product Partition Model

Fearnhead and Liu (2007) introduced their online algorithm for exact filtering of multiple changepoint problems called the *Direct Simulation* algorithm based on the previous MCMC methods proposed by Fearnhead (2006). Furthermore, they showed that the computational cost of this exact algorithm is quadratic in the number of observations, and therefore not suitable for online detection. In order to improve the performance of their system, they utilized resampling ideas from particle filters at the expense of introducing errors.

Direct Simulation Algorithm

Particle filters or Sequential Monte Carlo (SMC) methods (Gordon et al., 1993) are a class of stochastic sampling algorithms which allow approximation of a sequence of probability distributions and are used for estimating sequential Bayesian models. Particles (samples) are used to represent points in the distribution that is to be estimated and are assigned weights based on their approximate probabilities (Doucet et al., 2001). The number of particles can grow at each iteration or time step, and so some particles may need to be discarded. This necessitates the assignment of new weights to the remaining particles through a procedure called resampling (Mellor and Shapiro, 2013).

Particle Filters

One of the biggest advantages of the direct simulation method, over Gibbs sampler and reversible-jump MCMC, is that there is no need to ascertain whether the MCMC algorithm has converged or not. Moreover, MCMC techniques are far too computationally expensive for huge data sets and, hence, not desirable for online inference.

Xuan and Murphy (2007) applied the direct simulation algorithm in a multivariate setting, and evaluated the method on bee waggle dance dataset (Oh et al., 2006). Chopin (2007) also introduced a particle filtering algorithm for online and offline changepoint detection, but it is outperformed by Fearnhead and Liu’s method (Fearnhead and Liu, 2007).

Adams and MacKay (2007) proposed a generic approach with the aim of generating an accurate distribution of the next unseen datum in a sequence, given only data already observed, using a message passing algorithm in a recursive fashion. Their method was tested on three datasets: (1) coal-mining disasters, also studied as a retrospective problem by Raftery and Akman (1986); (2) daily returns of the Dow Jones Industrial Average, also studied by Hsu (1977) with a frequentist approach; and (3) nuclear magnetic response, also studied by Fearnhead (2006) using MCMC methods.

Adams and MacKay’s Method

They cast the mentioned product partition model into a Bayesian graphical model equivalent to a Hidden Markov Model (HMM) with a possibly infinite number of hidden states, as there can be as many change points as data observations (Paquet, 2007). An advantage of this setting is that the number of changepoints does not have to be specified in advance.

Similar to the work of Fearnhead and Liu (2007), their exact inference algorithm is not efficient, and has space and time requirements that grow linearly in time. Therefore, they suggest an approximate inference technique where run-lengths, the length of the segment between two consecutive changepoints, with assigned probability masses less than a threshold value are eliminated.

It is worth mentioning that Fearnhead and Liu’s direct simulation algorithm maintains a finite sample of the run-length distributions (by using particles), and so has the benefit of being certain on the upper bound of the algorithm’s space requirements (Mellor and Shapiro, 2013).

Since 2007, some researchers have expanded Adams and MacKay’s and Fearnhead and Liu’s work. For example, Wilson et al. (2010) have addressed one of the shortcomings of these algorithms: the assumption that the frequency with which changepoints occur, known as the *hazard rate*, is fixed and known in advance. They eliminated this restrictive assumption, and proposed a system that is also capable of learning the hazard rate in a recursive fashion. Caron et al. (2012) addressed another limitation: the need for knowledge of the static parameters of the model to infer the number of changepoints and their locations. They propose an extension of Fearnhead and Liu’s algorithm which allows them to estimate jointly these static parameters using a recursive maximum likelihood estimation strategy.

### 3.3.3 Other Bayesian Approaches

Some researchers (including Baxter and Oliver, 1996; Oliver et al., 1998; Viswanathan et al., 1999; and Fitzgibbon et al., 2002) have approached the changepoint detection problem as a time-series segmentation problem. In the segmentation problem, the data is partitioned into distinct homogeneous regions delimited by two consecutive changepoints. In these studies, the Minimum Message Length (MML) principle (Wallace, 2005) was utilized to address the segmentation problem. As MML is a powerful tool when dealing with large datasets, this approach has advantages in problems with long streams of data such as DNA sequences.

Minimum Message Length

# 4. Changepoint Detection on Twitter

Applying changepoint detection techniques for temporal analysis of tweets has been the subject of very few studies, which are discussed in this section. It is noteworthy that their changepoint detection methods are not suitable for our model as the first one rely on knowledge regarding the problem domain and the second one is an offline changepoint detection method.

Collier and Doan (2011) studied the tracking of infectious diseases on Twitter. In order to detect unexpected rises in the stream of messages for each of the syndromes they studied, they first classified tweets using both a Naïve Bayes Classifier and an SVM and then applied a changepoint detection algorithm called the Early Aberration and Reporting System (EARS) (Hutwagner et al., 2003), which reports an alert when its test value (number of tweets classified under a disease) exceeds a certain number of standard deviations above a historic mean. This method requires knowledge of the problem domain, which is a shortcoming of many simple statistical changepoint detection techniques, such as the famous CUSUM method[[4]](#footnote-4).

Liu et al. (2013) who carried out research closest to ours, developed a novel offline change detection algorithm called Relative Density-Ratio Estimation and evaluated their method, among other datasets, on the then publicly available CMU Twitter dataset, which is a set of tweets from February to October 2010. They tracked the degree of popularity of a topic by monitoring the frequency of some selected keywords. More specifically, they focused on events related to “Deepwater Horizon oil spill in the Gulf of Mexico” which occurred on April 20, 2010. They used the frequencies of 10 hand-selected keywords (Figure 1), then performed changepoint detection directly on the 10-dimensional data to capture correlation changes between multiple keywords, in addition to changes in the frequency of each keyword. For evaluation, they referred to the Wikipedia entry “Timeline of the Deepwater Horizon oil spill”[[5]](#footnote-5) as a real-world event source and matched the notable updates of the news story to the changepoints in their model (Figure 2). We will take a similar approach in our evaluation.

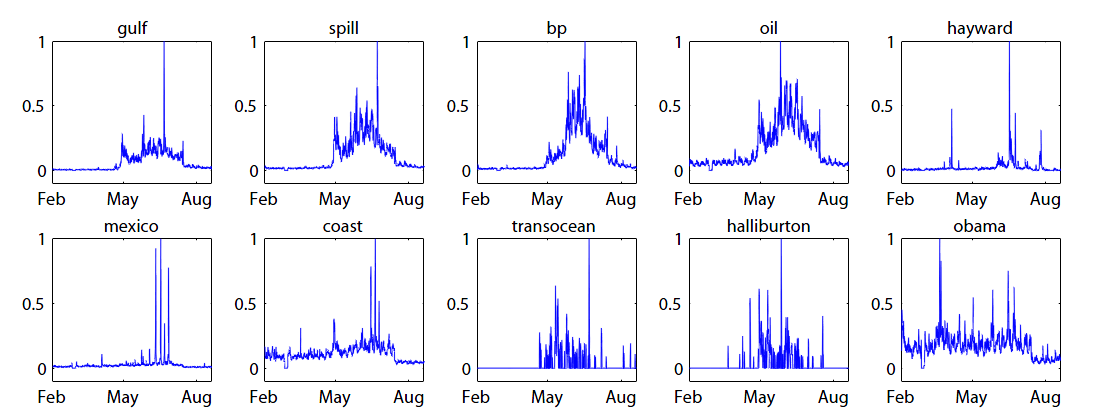
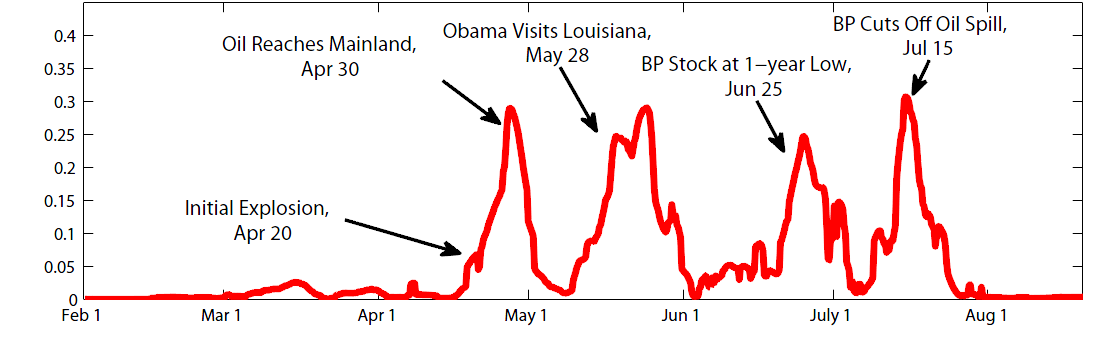


Figure 2 Change-point score obtained by Liu et al. (2013) is plotted and the four occurrences of important real-world events show the development of this news story

Figure 2 Normalized frequencies of the ten chosen keywords

# References:

Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Analysing user modelling on twitter for personalized news recommendations. In *User Modelling, Adaption and Personalization* (pp. 1-12). Springer Berlin Heidelberg.

About.twitter.com. (2014). About Twitter, Inc. | About. Retrieved 5 November 2014, from https://about.twitter.com/company.

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection.*arXiv preprint arXiv:0710.3742*.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, *50*(1-2), 5-43.

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2011). Spatio-temporal analysis of topic popularity in twitter. *arXiv preprint arXiv:1111.2904*.

Bansal, N., & Koudas, N. (2007). Blogscope: a system for online analysis of high volume text streams. In *Proceedings of the 33rd international conference on Very large data bases* (pp. 1410-1413). VLDB Endowment.

Barlow, J. S., Creutzfeldt, O. D., Michael, D., Houchin, J., & Epelbaum, H. (1981). Automatic adaptive segmentation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology*, *51*(5), 512-525.

Barry, D., & Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, 260-279.

Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application* (Vol. 104). Englewood Cliffs: Prentice Hall.

Baxter, R. A., & Oliver, J. J. (1996). The kindest cut: minimum message length segmentation. In *Algorithmic Learning Theory* (pp. 83-90). Springer Berlin Heidelberg.

Beaulieu, C., Chen, J., & Sarmiento, J. L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *370*(1962), 1228-1249.

Bodenstein, G., & Praetorius, H. M. (1977). Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*,*65*(5), 642-652.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Bovolo, F., Bruzzone, L., & Marconcini, M. (2008). A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. *Geoscience and Remote Sensing, IEEE Transactions on*, *46*(7), 2070-2082.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1-10). IEEE.

Brodsky, E., & Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems* (No. 243). Springer.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied statistics*, 389-405.

Cataldi, M., Di Caro, L., & Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (p. 4). ACM.

Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of econometrics*, *86*(2), 221-241.

Chopin, N. (2007). Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, *59*(2), 349-366.

Collier, N., & Doan, S. (2011). Syndromic classification of twitter messages.*arXiv preprint arXiv:1110.3094*.

Csörgö, M., & Horváth, L. (1997). *Limit theorems in change-point analysis*. New York: Wiley.

Denison, D. G. T., & Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 143-149.

Doucet, A., De Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential Monte Carlo methods in practice*. Springer.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, *16*(2), 203-213.

Fearnhead, P., & Liu, Z. (2007). On‐line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 589-605.

Fitzgibbon, L. J., Dowe, D. L., & Allison, L. (2002). Change-point estimation using new minimum message length approximations. In *PRICAI 2002: Trends in Artificial Intelligence* (pp. 244-254). Springer Berlin Heidelberg.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6), 721-741.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)* (Vol. 140, No. 2, pp. 107-113). IET Digital Library.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711-732.

Gustafsson, F., & Gustafsson, F. (2000). *Adaptive filtering and change detection* (Vol. 1). New York: Wiley.

Habib, T., Inglada, J., Mercier, G., & Chanussot, J. (2009). Support vector reduction in SVM algorithm for abrupt change detection in remote sensing. *Geoscience and Remote Sensing Letters, IEEE*, *6*(3), 606-610.

Hsu, D. A. (1977). Tests for variance shift at an unknown time point. *Applied Statistics*, 279-284.

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (pp. 173-178). ACM.

Hutwagner, M. L., Thompson, M. W., Seeman, G. M., & Treadwell, T. (2003). The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health*, *80*(1), i89-i96.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, *60*(11), 2169-2188.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65). ACM.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110-125.

Koop, G. M., & Potter, S. (2004). Forecasting and estimating multiple change-point models with an unknown number of change points.

Kramer, A. D. (2010). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 287-290). ACM.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79-86.

Kunneman, F. A., Liebrecht, C. C., & van den Bosch, A. P. J. (2014). The (Un) Predictability of Emotional Hashtags in Twitter.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World Wide Web* (pp. 591-600). ACM.

Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on* (pp. 411-416). IEEE.

Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector-tracking epidemics on Twitter. In *Machine Learning and Knowledge Discovery in Databases* (pp. 599-602). Springer Berlin Heidelberg.

Lampos, V., Preotiuc-Pietro, D., & Cohn, T. (2013). A user-centric model of voting intention from Social Media. In *ACL (1)* (pp. 993-1003).

Lim, Kar Wai, and Wray Buntine. "Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon." (2014).

Liu, S., Yamada, M., Collier, N., & Sugiyama, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, *43*, 72-83.

Madsen, H. (2007). *Time series analysis*. Boca Raton, Florida: Chapman & Hall/CRC Press.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labelling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889-892). ACM.

Mellor, J., & Shapiro, J. (2013). Thompson Sampling in Switching Environments with Bayesian Online Change Detection. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (pp. 442-450).

Naaman, M., Boase, J., & Lai, C. H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 189-192). ACM.

Oh, S. M., Rehg, J. M., & Dellaert, F. (2006). Parameterized duration modelling for switching linear dynamic systems. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 2, pp. 1694-1700). IEEE.

Oliver, J. J., Baxter, R. A., & Wallace, C. S. (1998). Minimum message length segmentation. In *Research and Development in Knowledge Discovery and Data Mining* (pp. 222-233). Springer Berlin Heidelberg.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 100-115.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*(pp. 79-86). Association for Computational Linguistics.

Paquet, U. (2007). Empirical Bayesian change point detection. *Graphical Models*, *1995*, 1-20.

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. In *ICWSM* (pp. 265-272).

Preotiuc-Pietro, D., & Cohn, T. (2013). A temporal model of text periodicities using Gaussian Processes. In *EMNLP* (pp. 977-988).

Raftery, A. E., & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 85-89.

Reeves, J., Chen, J., Wang, X. L., Lund, R., & Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, *46*(6), 900-915.

Ritter, A., Clark, S., & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524-1534). Association for Computational Linguistics.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 695-704). ACM.

Roth, T., Sprau, P., Naguib, M., & Amrhein, V. (2012). Sexually selected signaling in birds: a case for Bayesian change-point analysis of behavioral routines. *The Auk*, *129*(4), 660-669.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web* (pp. 851-860). ACM.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, 159-178.

Takahashi, T., Tomioka, R., & Yamanishi, K. (2011). Discovering emerging topics in social streams via link anomaly detection. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 1230-1235). IEEE.

Thelwall, M., & Prabowo, R. (2007). Identifying and characterizing public science‐related fears from RSS feeds. *Journal of the American Society for Information Science and Technology*, *58*(3), 379-390.

Viswanathan, M., Wallace, C. S., Dowe, D. L., & Korb, K. B. (1999). Finding Outpoints in Noisy Binary Sequences—A Revised Empirical Evaluation. In*Advanced Topics in Artificial Intelligence* (pp. 405-416). Springer Berlin Heidelberg.

Wallace, C. S. (2005). *Statistical and inductive inference by minimum message length*. New York: Springer.

Wang, Y., Wu, C., Ji, Z., Wang, B., & Liang, Y. (2011). Non-parametric change-point method for differential gene expression detection. *PloS one*, *6*(5), e20060.

Whiteley, N., Andrieu, C., & Doucet, A. (2011). Bayesian computational methods for inference in multiple change-points models.

Wigand, F. D. L. (2010). Twitter in government: Building relationships one tweet at a time. In *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on* (pp. 563-567). IEEE.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2010). Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, *22*(9), 2452-2476.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011a). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 705-714). ACM.

Wu, S., Tan, C., Kleinberg, J. M., & Macy, M. W. (2011b). Does Bad News Go Away Faster?. In *ICWSM*.

Xuan, X., & Murphy, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning* (pp. 1055-1062). ACM.

Yamanishi, K., & Takeuchi, J. I. (2002). A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 676-681). ACM.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 177-186). ACM.

1. Emoji are much like emoticons, but with a wider range as they are not restricted to ASCII characters. [↑](#footnote-ref-1)
2. Hashtags are user-generated labels included in online posts by their authors to categorize the post under a topic or make it part of a conversation. This metadata tag is in the form of a word or an unspaced phrase prefixed with the "#" character. [↑](#footnote-ref-2)
3. Other Twitter users' names preceded by an @ character. [↑](#footnote-ref-3)
4. CUSUM, in its simple form, calculates the cumulative sum of the data points and identifies a change if this sum exceeds a threshold value (see Page (1954) for a more complete description of the method and Basseville and Nikiforov (1993) for some of the variations applied to the original method). [↑](#footnote-ref-4)
5. <http://en.wikipedia.org/wiki/Timeline_of_the_Deepwater_Horizon_oil_spill> [↑](#footnote-ref-5)